NASA INTERNATIONAL SPACE APPS CHALLENGE

Challenge

CAN AI PRESERVE OUR SCIENCE LEGACY?

NASA

AI SCIENCE KNOWLEDGE (AISK NASA)

Isabella Llinás; Christian Duarte; Edward Camelo

UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

The Challenge

The NASA Technical Report Server (NTRS) includes hundreds of thousands of items containing scientific and technical information (STI) created or funded by NASA. Imagine how difficult it can be to locate desired information in such a large repository! Your challenge is to develop a technique using Artificial Intelligence (AI) to improve the accessibility and discoverability of records in the public NTRS.

UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

NASA

NASA INTERNATIONAL SPACE APPS CHALLENGE

First Step: Extract the data

API CONECTION

Direct Connection
Server connection
Hosting connection

NTRS

Api for connect to the database

Getting the data

https

NTRS

Getting the data (scrap python)

aws

Azure

In a Bucket and k3s Save the pdf files

A storage of objects

NPL
Python's libraries

for *.pdf:

1. Read the pdf
2. Text analitic
   - Keywords
   - Definition
3. Summary

Get a summary with NPL

docker

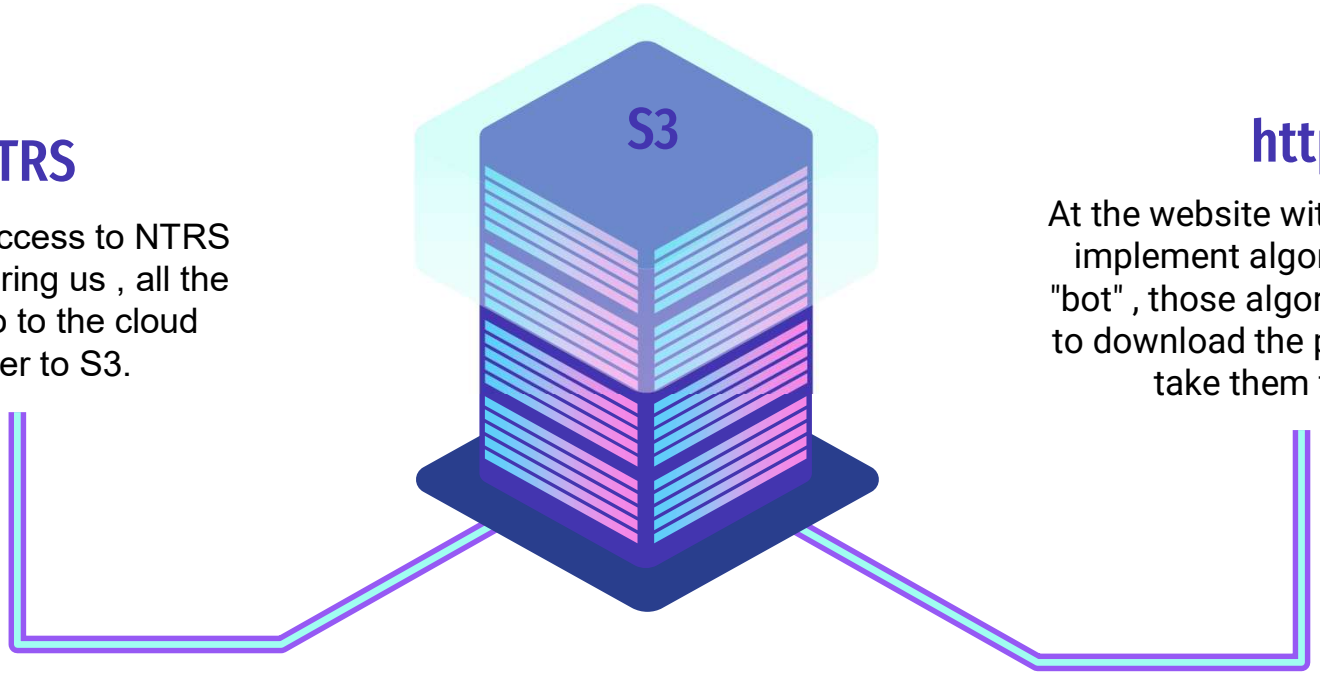
UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

First Step: Extract the data

NTRS

If we have access to NTRS that NASA bring us , all the pdf files go to the cloud and later to S3.



https

At the website with the aim to implement algorithms of a "bot" , those algorithms going to download the pdf files and take them to s3

No matter the process the files will arrive to S3 where using libraries of python and an algorithm of reading, it is going to look the keywords and for each pdf file will have a summary.

UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

Scrapping Python (AI)

Scrapping Python

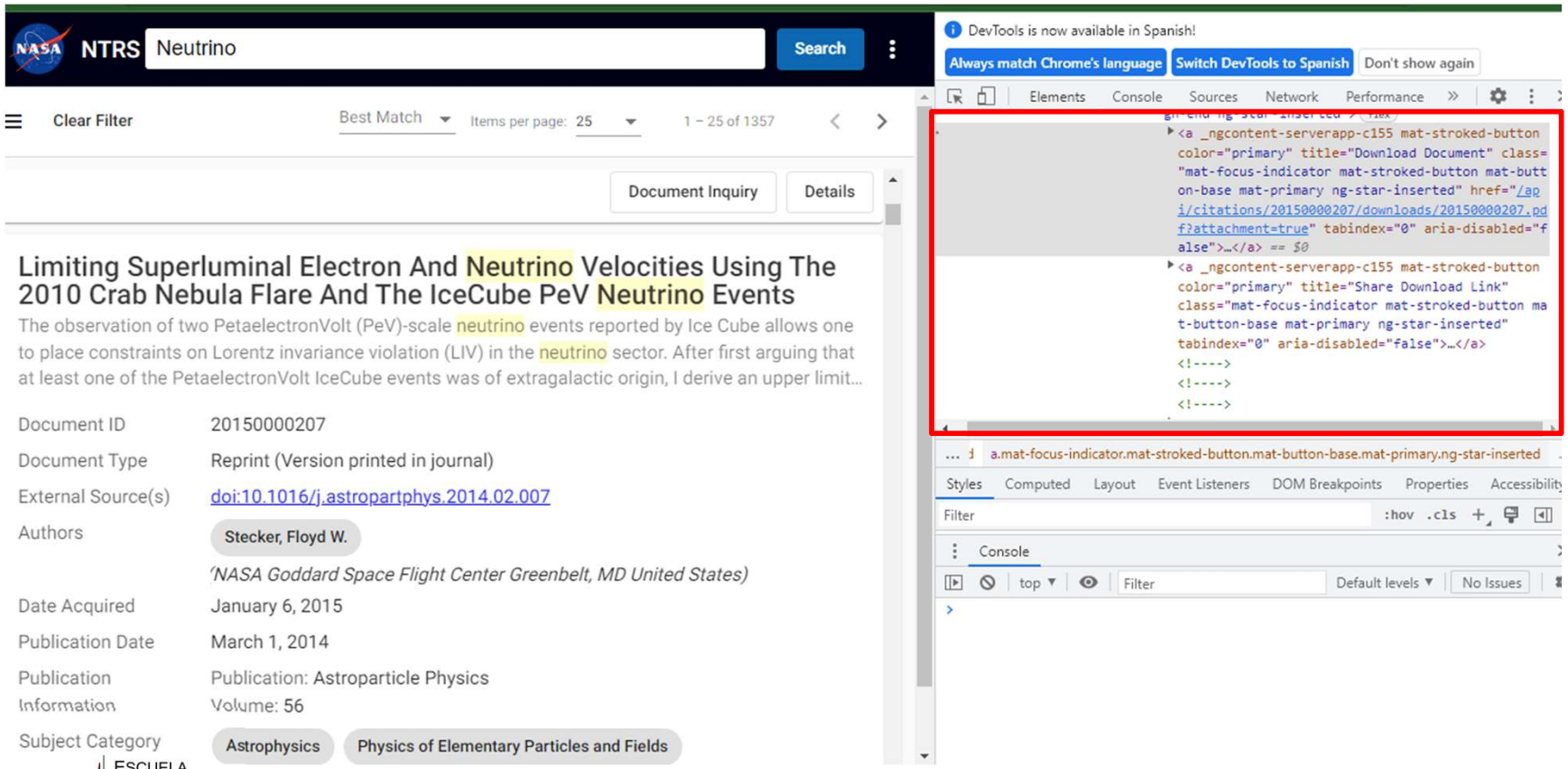
Web scraping is the process of collecting and parsing raw data from the Web, and the Python community has come up with some pretty powerful web scraping tools.

UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

NASA

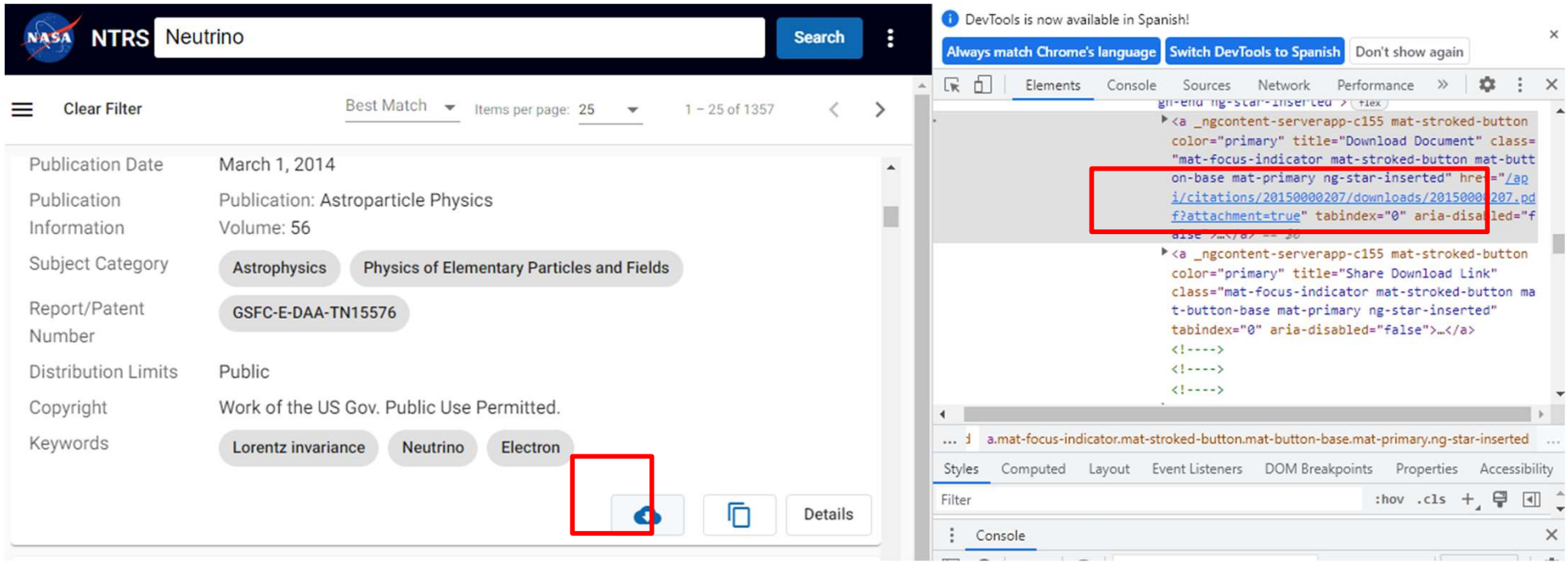
Scrapping



UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

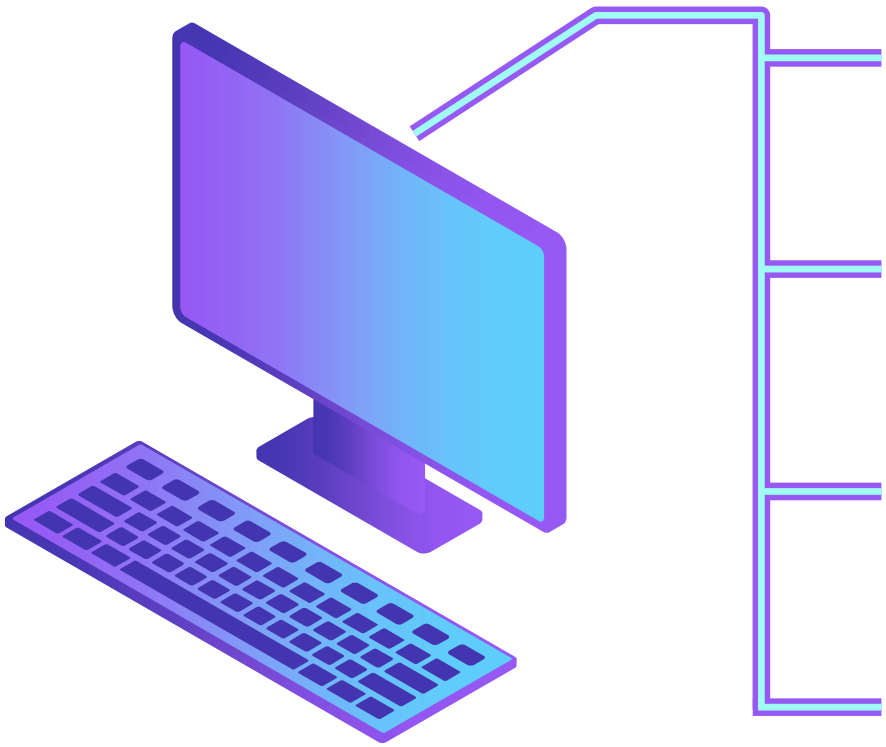
Scrapping



UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

What is a container and how do I use it?



What is a container?

Are a solution to the problem of how to get software to run reliably when moved from one computing environment to another.

Why is important get a container?

Pretend You're going to test using Python 2.7, and then it's going to run on Python 3 in production and something weird will happen and there is no solution for that problem, then the job does not work.

How a container is a solution?

A container consist of an entire runtime environment: an application, plus all its dependencies, libraries and other binaries, and configuration files needed to run it, bundled into one package so if the job need anything else to run, then the container will provide it.

How use a container?

To use a container we have to create a doocker file and use it in our pdf files and the container starts to work

UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

NASA

NPL Python's libraries

Natural Language Analysis with Python NLTK

NLTK (Natural Language Toolkit)

It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Definition

is a subfield of artificial intelligence (AI). It helps machines process and understand the human language so that they can automatically perform repetitive tasks

Spacy

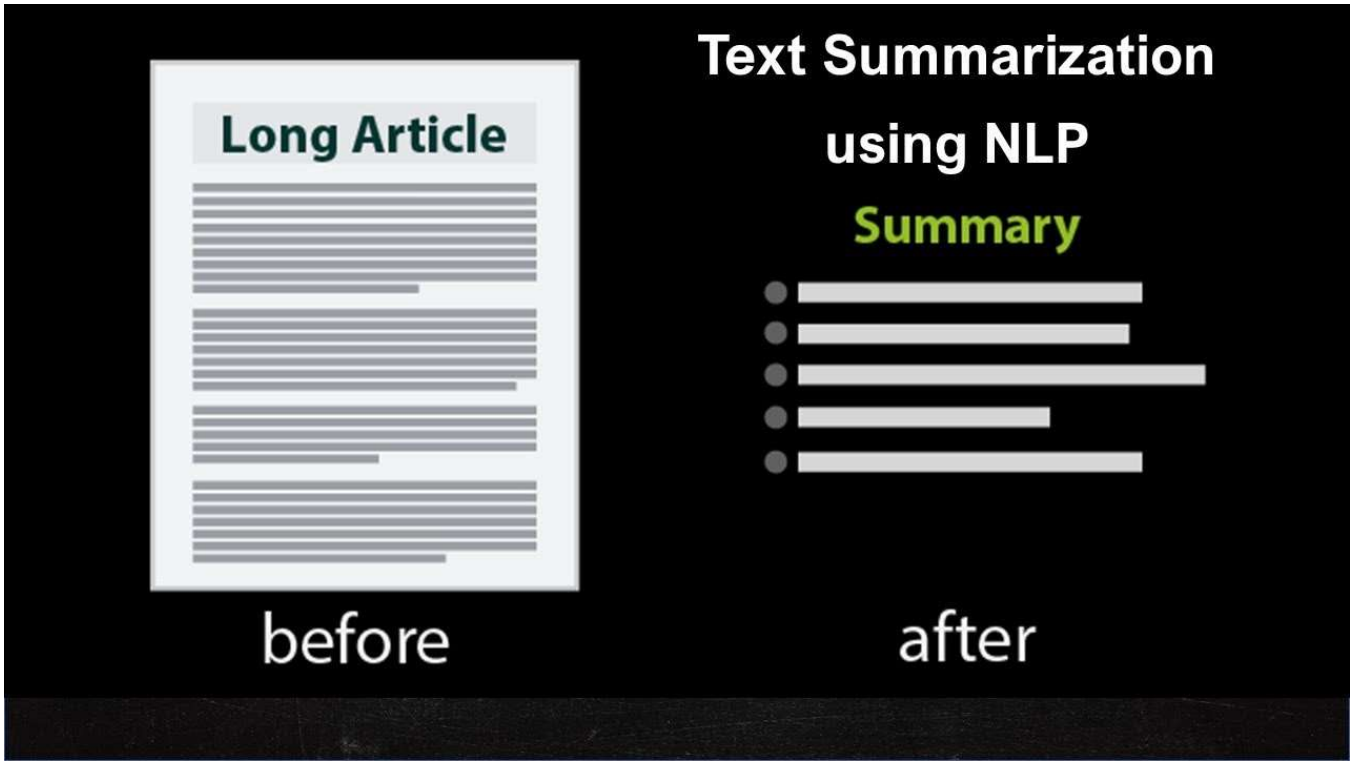
Is the main competit or of the NTLK. Both libraries have the same function

Python libraries of IA interaction.

UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

NASA



Natural Language Analysis
with Python NLTK

UNIVERSIDAD
SERGIO ARBOLEDA

ESCUELA
COLOMBIANA
DE INGENIERÍA
JULIO GARAVITO

NASA

Second step: Create a database

id_article	name_article	summary_article

Article table

User table

id_user	name_user	univesity	Gender	interest_user	knowledge_user	orcid_user	CVLAC_user

id_interaction	id_user	date	timestap	id_article	ranking

interactions table

id_article	id_user	Rating_recomendation

Recommendations table

UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

Second step: Create a database

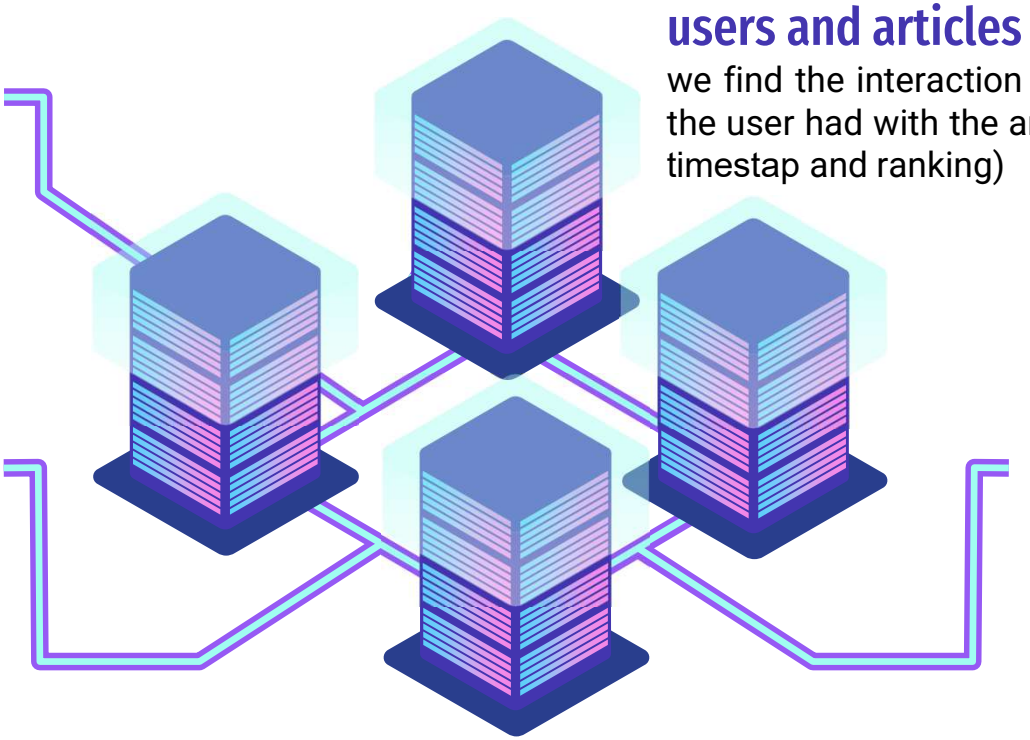
Text processed RESUME

Articles

Where we find the information about the id, name, and summary of each article

Interactions between the users and articles

we find the interaction information that the user had with the article (user, date, timestap and ranking)



Recommendations

We find the information about the id of the article, id of the user, and percent of interest

Users

We find the personal information of the user like his id, name. University, genere, interest, knowledge, CVLac of the user.

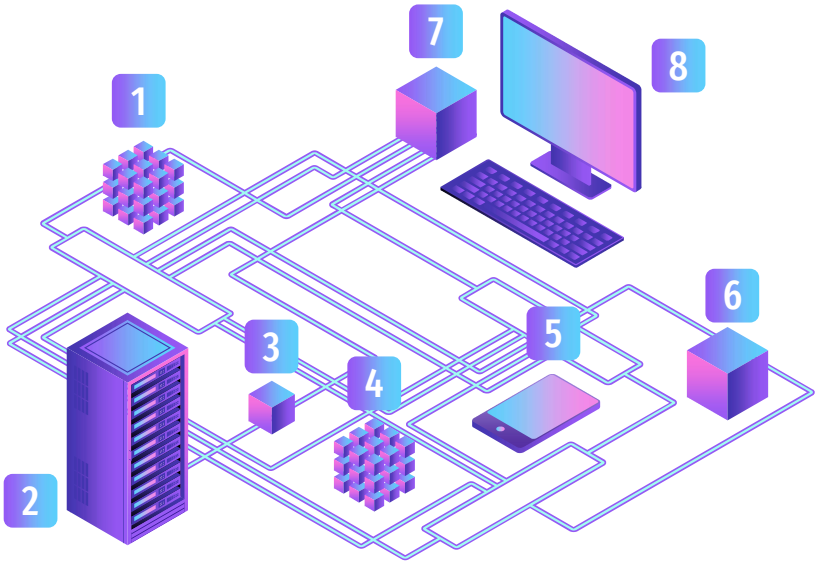
UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

The new database is being consumed by the website

the first table (articles) is completed when all the articles have been obtained the first time

the second table (users) is completed with the information that the user provides on the web site



the third table (interactions) is completed thanks to all the interaction that the user has with the article from the time and type of interaction

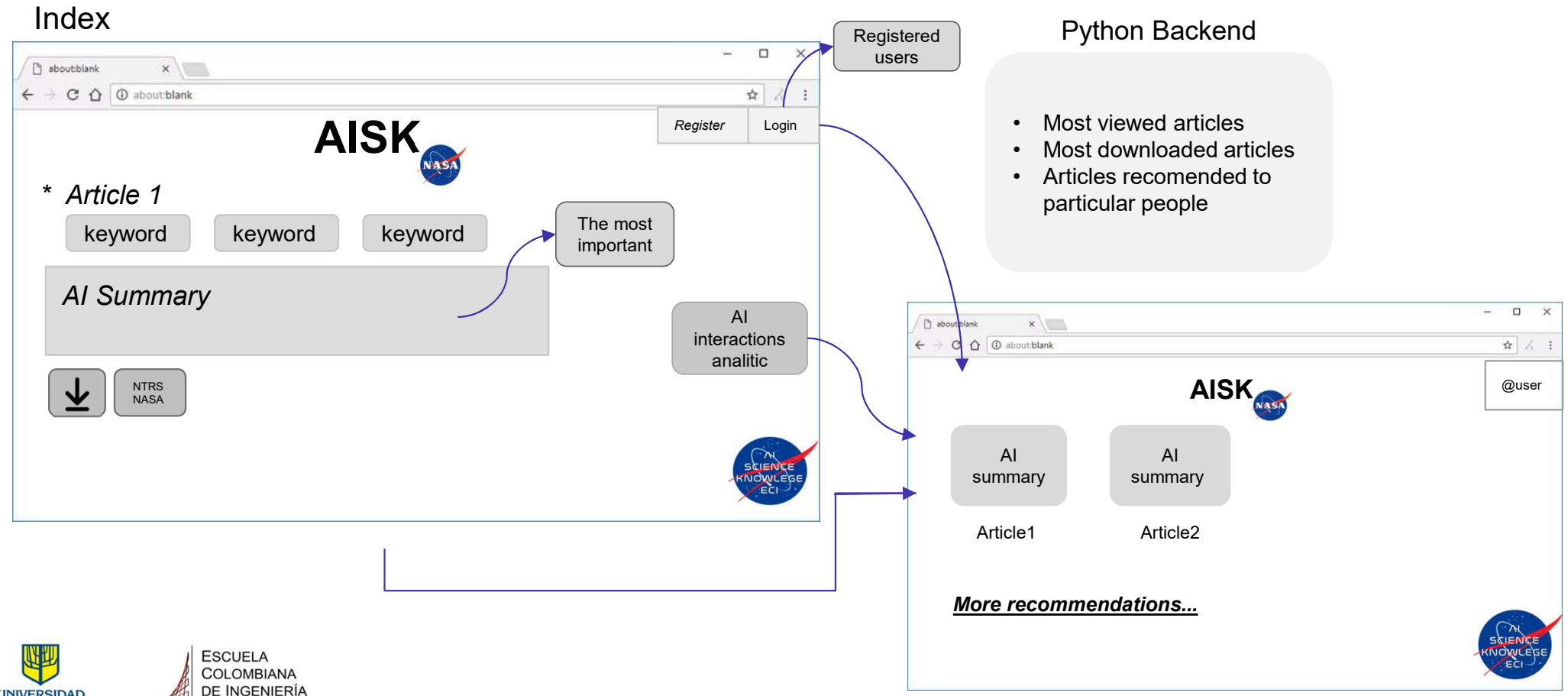
the last table (recommendations) is completed thanks to the interactions of each of the users showing most viewed articles, most downloaded articles

UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

NASA

third step: Create a website



UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

Algorithms for items recommendations

recommends an item to a user if similar users liked this item before.

surprise

Provide various ready-to-use prediction algorithms such as baseline algorithms, neighborhood methods, matrix factorization-based (SVD, PMF, SVD++, NMF), and many others. Also, various similarity measures (cosine, MSD, pearson...) are built-in.

from surprise import SVD++

To get the access under level of hardware like audio, keyboard, mouse, and other devices

from surprise import Dataset

Is the basic data container in PyMVPA. It serves as the primary form of data storage

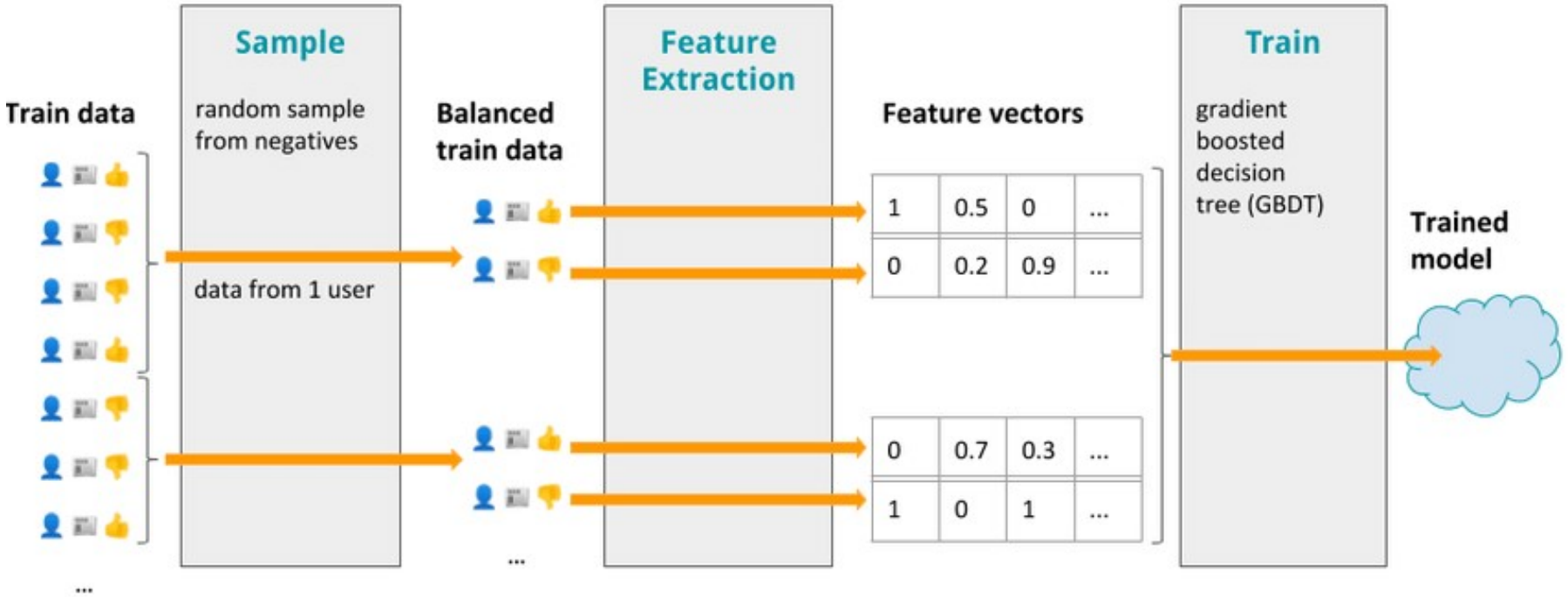
from surprise import accuracy

Is the percentage of data that are correctly classified, which ranges from 0 to 1.

UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

How works the pipeline recommenders



UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

What is Docker and why do we use it?

Docker

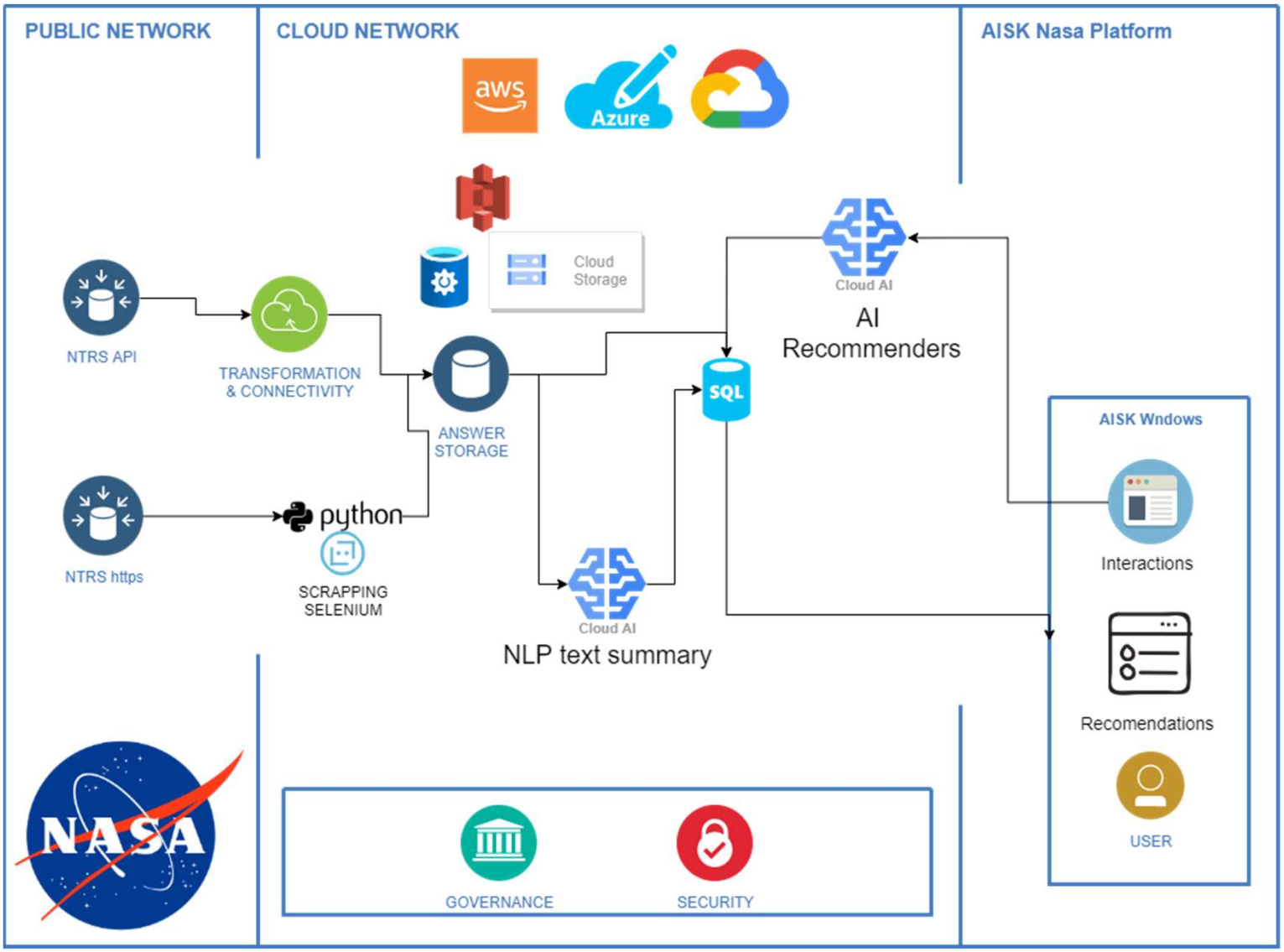
Docker is a software platform that allows you to build, test, and deploy applications quickly. Docker packages software into standardized units called containers that have everything the software needs to run including libraries, system tools, code, and runtime.

UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

NASA

AISK

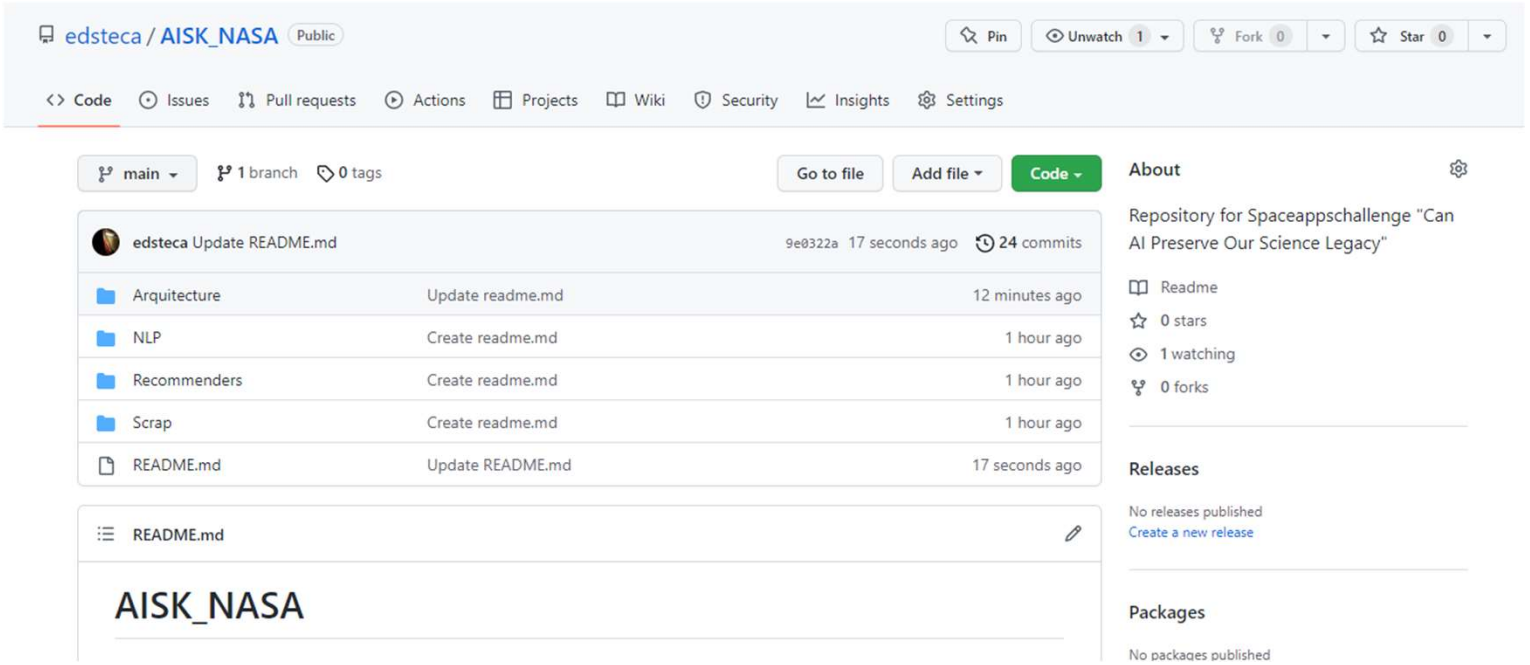


UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

GitHub Repository

https://github.com/edsteca/AISK_NASA



UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

NASA

References

Comprehensive Guide to Text Summarization using Deep Learning in Python
https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/

- Recommenders https://github.com/microsoft/recommenders
- NLP examples https://github.com/microsoft/nlp-recipes
- Databases http://www.pymvpa.org/tutorial_datasets.html
- Recomendaciones usando Python https://ourcodeworld.co/articulos/leer/1101/construyendo-un-sistema-de-recomendacion-simple-en-python
- Read pdf file using Python https://www.delftstack.com/es/howto/python/read-pdf-in-python/#:~:text=PDFplumber%20es%20un%20m%C3%B3dulo%20de,para%20leer%20un%20archivo%20PDF.
- What is S3 https://aws.amazon.com/es/s3/getting-started/
-What is a container https://www.cio.com/article/247005/what-are-containers-and-why-do-you-need-them.html
- Text Summarizer https://github.com/aravindpai/How-to-build-own-text-summarizer-using-deep-learning/blob/master/How_to_build_own_text_summarizer_using_deep_learning.ipynb

UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

NASA

NASA

NASA INTERNATIONAL SPACE APPS CHALLENGE

UNIVERSIDAD SERGIO ARBOLEDA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO